Profiling Speech Sound Disorders for Clinical Validation of the Computer Articulation Instrument

van Haaften, Leenke; Diepeveen, Sanne; Terband, Hayo; Vermeij, Bernadette; van den Engel-Hoek, Lenie; de Swart, Bert; Maassen, Ben

Published in:
American journal of speech-language pathology

DOI:
10.1044/2018_AJSLP-MSC18-18-0112

AJSLP

Research Article

Profiling Speech Sound Disorders for Clinical Validation of the Computer Articulation Instrument

Leenke van Haaften,[a] Sanne Diepeveen,[a,b] Hayo Terband,[c] Bernadette Vermeij,[d] Lenie van den Engel-Hoek,[a] Bert de Swart,[a,b] and Ben Maassen[e]

Purpose: The current article presents data from 2 studies on clinical groups of children referred for speech assessment. The aims of these studies are to validate the Computer Articulation Instrument (CAI) with the known-group validation method and to determine the differential diagnostic power of the resulting speech profiles.
Method: Study 1 examined known-group validity by comparing the scores of 93 children diagnosed with speech-language difficulties on the picture naming (PN) task of the CAI with intelligibility judgments given by speech-language pathologists. In Study 2, the speech profiles of 41 children diagnosed with speech sound disorders (SSDs), consisting of 4–6 factor scores extracted from the 4 tasks of the CAI, namely, PN, nonword imitation (NWI), word and nonword repetition, and maximum repetition rate (MRR), were validated against clinical judgments of severity of the SSD given by speech-language pathologists.
Results: In Study 1, a repeated-measures analysis of variance revealed a significant effect of intelligibility level on the PN performance of the CAI and there were highly significant correlations between intelligibility and PN performance in the expected direction. Neither intelligibility level nor PN performance was related to nonverbal intelligence and language scores. The analysis of variance and a series of *t* tests in Study 2 revealed significant differences between the moderate and severe groups for the CAI factors based on PN and NWI and the bisyllabic and trisyllabic sequences of MRR, but not for the factor word and nonword proportion of whole-word variability based on word and nonword repetition, and the monosyllabic sequences of MRR. These results suggest that, especially, the tasks PN, NWI, and the bisyllabic and trisyllabic sequences of MRR are most sensitive for diagnosing SSDs.
Conclusions: The findings of these 2 studies support the known-group validity of the CAI. Together with the results of a previous study of our group on reliability and validity (van Haaften et al., 2019), we can conclude that the CAI is a reliable and valid tool for assessment of children with SSDs.

Children with speech production problems are one of the four subtypes that can be distinguished in children with a specific language impairment (Van Weerdenburg, Verhoeven, & Van Balkom, 2006). They show a specific profile as compared to the other subtypes of children with language impairments: difficulties with lexical–semantic abilities, with auditory conceptualization, or with verbal sequential memory (Van Weerdenburg et al., 2006). Recently, Bishop et al. (2017) proposed to use the term *developmental language disorder* (DLD) when a language disorder was not associated with a known biomedical etiology. They state that DLD is a heterogeneous category that encompasses a wide range of problems, including expressive phonological problems. Phonological problems in preschoolers that are not accompanied by other language problems do not meet the criteria for DLD. Therefore, Bishop et al. propose to use the more general term *speech*

[a]Department of Rehabilitation, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, the Netherlands
[b]HAN University of Applied Sciences, Nijmegen, the Netherlands
[c]Utrecht Institute of Linguistics-OTS, Utrecht University, the Netherlands
[d]Dutch Foundation for the Deaf and Hard of Hearing Child, Amsterdam, the Netherlands
[e]Center for Language and Cognition, Groningen University, the Netherlands

Correspondence to Ben Maassen: b.a.m.maassen@rug.nl

Editor-in-Chief: Kristie Spencer
Editor: Kirrie Ballard

Received May 18, 2018
Revision received October 11, 2018
Accepted November 12, 2018
https://doi.org/10.1044/2018_AJSLP-MSC18-18-0112

Publisher Note: This article is part of the Special Issue: Selected Papers From the 2018 Conference on Motor Speech—Clinical Science and Innovations.

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

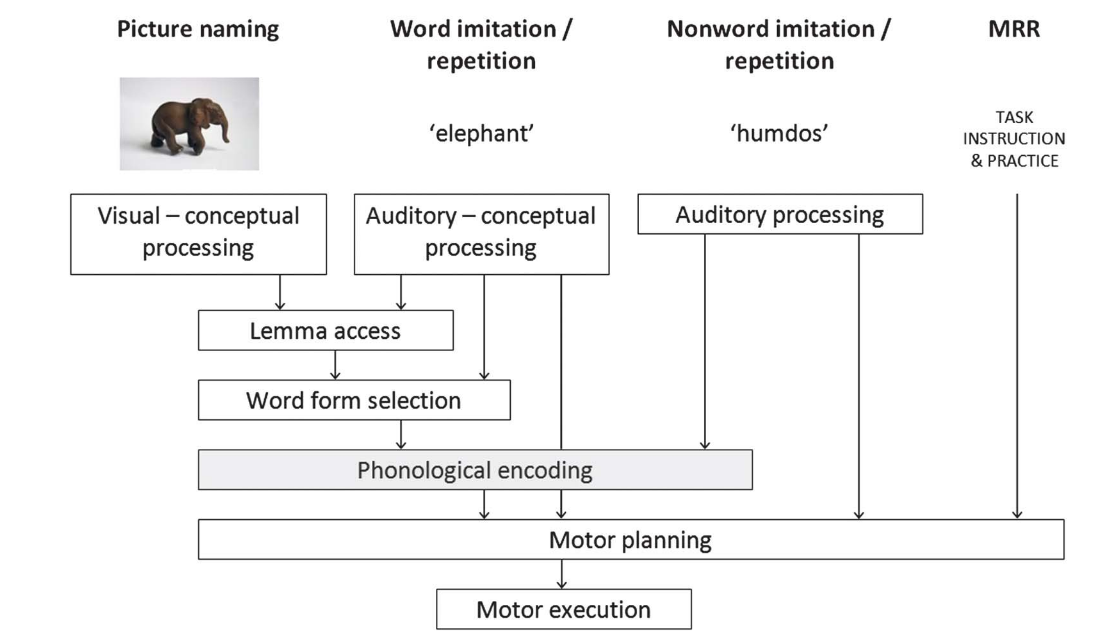
sound disorder (SSD) for such cases. SSD is an umbrella term that includes expressive phonological problems and problems with speech production that have motor or physical origins or involve misarticulations such as a lisp, where a sound is produced in a distorted way without losing the contrast with other sounds. Children with SSDs are one of the most common clinical populations for speech-language pathologists (SLPs; Mullen & Schooling, 2010); the reported prevalence is highly variable, ranging from 2.3% to 24.6% (Eadie et al., 2015; Law, Boyle, Harris, Harkness, & Nye, 2000). They form a heterogeneous group, showing variability in severity, etiology, proximal causes, speech error characteristics, and response to treatment (Dodd, 2011).

There are several widely recognized classification systems for SSDs featuring a variety of approaches, namely, etiology, descriptive linguistics, and psycholinguistic and psychomotor processing (Waring & Knight, 2013). In current practice, symptom patterns form the basis of diagnostic classification (Dodd, 1995b, 2014). The Speech Disorders Classification System described by Shriberg et al. (2017) divides SSDs into three classes, based on etiology: speech delay, speech errors, and motor speech disorder (MSD; including dysarthria, childhood apraxia of speech [CAS], and MSD–not otherwise specified). Examples of symptoms of MSD include slow speech rate, distorted substitutions of speech sounds, increased difficulty with multisyllabic words, and prosodic errors. Yet, there is no validated list of diagnostic patterns for differential diagnosis of SSDs. For example, one of the speech symptoms that is described for different types of SSDs is inconsistency of speech errors. From a phonological point of view, high inconsistency of speech errors could indicate an unstable phonological system, also called a *phonological planning deficit* (Dodd, 1995a; Macrae, Tyler, & Lewis, 2014), or unstable lexical representations (Sosa & Stoel-Gammon, 2012). However, inconsistency is also a characteristic of CAS (Davis, Jakielski, & Marquardt, 1998; Forrest, 2003; Iuzzini-Seigel, Hogan, & Green, 2017). In the latter case, inconsistency is explained by an unstable motor system (articulomotor planning and programming). Thus, the same symptom can refer to different underlying deficits, and the same deficit can result in different symptoms, leading to a wide variety of symptoms within subtypes and much symptomatic overlap between subtypes of SSDs. Therefore, in clinical practice, a reorientation from behavioral diagnostics to process-oriented diagnostics is required in order to reveal the proximal causes of SSDs (Terband & Maassen, 2012).

Psycholinguistic and psychomotor models give a conceptual basis to analyze speech disorders and form the basis for a process-oriented diagnostic classification system based on the identification of the breakdown in the chain of sequential and parallel speech processes (Baker, Croot, McLeod, & Paul, 2001). Rather than categorization of SSDs based on single symptoms or sets of symptoms, process-oriented diagnostics primarily focus on speech profiles comprising clustered symptoms that can be interpreted in terms of the underlying speech production processes. An example of a psycholinguistic processing model is the model described by Levelt (1989), in which "conceptualizing a preverbal message," either from memory or from perception, is the first process in speaking. The next process is formulating a word or sentence, driven by two steps of lexicalization: selecting a lemma, containing meaning and grammatical information, and the corresponding lexeme or word form, which forms the input for the next stage of phonological encoding. *Phonological encoding* entails specifying the sequence of speech sounds together with their syllabic and prosodic structure. These syllables are the basic units of articulomotor planning and programming. The final process of actually performing the articulatory movements is *execution*, resulting in an acoustic speech signal (Maassen & Terband, 2015). Levelt, Roelofs, and Meyer (1999) validated this processing model with normal speech production data, and Nijland (2003) further elaborated on the planning, execution, and monitoring stages of the model and applied it to analyses of SSDs. By conducting different speech experiments in children with CAS, Nijland could conclude that both phonetic planning and motor programming are deviant in children with CAS. Levelt's model is relevant for analyzing SSDs because of the stages lexeme retrieval, phonological encoding, and self-monitoring, which are the processes underlying consistent and inconsistent phonological disorder (PD). MSDs, of which CAS and dysarthria are the main diagnostic categories, can be described by means of the motor planning, programming, and execution processes. However, the main objective of a process-oriented approach is not to categorize but to give a complete characterization of the speech profile, such that underlying processing deficits can be identified. Insight into the deficits that might be the underlying causes of the child's difficulty requires an extensive analysis of a child's performance on a range of speech tasks that reflect different levels of processing. Based on these premises, the Computer Articulation Instrument (CAI) was developed (Maassen et al., 2019). The CAI consists of a battery of speech production tasks and is based on a series of studies of Dutch children with developmental and acquired SSDs (Nijland, Maassen, & van der Meulen, 2003; Nijland, Maassen, van der Meulen, Gabreëls, et al., 2003; Nijland, Terband, & Maassen, 2015; Thoonen, Maassen, Gabreels, & Schreuder, 1999; Thoonen, Maassen, Gabreëls, & Schreuder, 1994). The CAI has a modular structure and provides an interactive administration and scoring of four speech tasks. The tasks comprise (a) picture naming (PN), (b) nonword imitation (NWI), (c) word and nonword repetition (WR and NWR), and (d) maximum repetition rate (MRR), thereby covering phonological and speech motor skills.

As demonstrated in Figure 1, PN taps into the whole chain of speech processes, from preverbal visual–conceptual processing to lemma access, word-form selection, phonological encoding, motor planning, and articulation (motor execution; Maassen & Terband, 2015). During NWI, a child is asked to reproduce nonwords (or nonsense words). In contrast to PN, a child cannot revert to its lexicon during this task, and thus the child either needs to analyze the phonological structure of the nonword directly, addressing the phonological decoding and encoding system, or follows the

Figure 1. The speech production processes assessed in the four tasks of the Computer Articulation Instrument (Maassen & Terband, 2015; Figure 15.2). MRR = maximum repetition rate.



auditory-to-motor-planning pathway. In WR and NWR, a child is asked to repeat a word or nonword five times. This task aims to assess variability in speech production, which occurs when a child uses multiple productions of the same word or nonword. MRR is a pure motor task (articulomotor planning and programming) and does not require any knowledge of words, syllables, or phonemes. The evaluation of speech production in the CAI is based on phonetic transcriptions and acoustic measurements. Both the tasks and speech analyses are computer implemented (van Haaften et al., 2019). Rather than focusing on single diagnostic markers, two types of analyses are conducted within the CAI: (a) objective and quantitative assessment of symptoms and (b) contrasting severity of symptoms across tasks. The outcome of this assessment battery is a speech performance profile that can be interpreted as characteristics of breakdown in underlying processes. Normative data from 1,524 children in the age range of 2;0–6;11 (years;months) have been collected, such that performance on the CAI as a whole, as well as the profile of performances on the different tasks, can be quantified in percentile scores, which allows for interpretation in terms of strengths and weaknesses (Maassen et al., 2019).

In a previous study of our research group, we assessed the psychometric properties of the CAI, including reliability and construct validity (van Haaften et al., 2019). Overall, sufficient to good values were found for interrater reliability, but intraclass correlation coefficients on test–retest reliability were low, probably due to better performance at retest reflecting a test–retest learning effect in addition to normal development. The study also described two aspects of construct validity. The first aspect, criterion validity, was confirmed by clear and significant age trends in CAI parameters in a large sample of typically developing children aged between 2 and 7 years. The second aspect of construct validity, structural validity, was assessed by factor analysis and correlations. Factor analyses on a total number of 20 parameters revealed five meaningful factors: PN; segmental quality of NWI (NWI-Seg); quality of syllabic structure of NWI (NWI-Syll); word and nonword proportion of whole-word variability (PWV), based on WR and NWR; and MRR. Weak correlations were found between CAI factor scores, indicating the independent contribution of each factor to the speech profile.

Further steps are needed in the validation process of the CAI. The ultimate goal is to assess the strengths of the five CAI factors in identifying the breakdown of speech processes in children with SSDs (process-oriented diagnostics), which will be described in future articles. The more immediate step, determining known-group validity, is presented in the current study. Known-group validity is a third aspect of construct validity and refers to the degree to which a measure is sensitive to differentiate between subgroups that are hypothesized to have different scores (Portney & Watkins, 2009). To assess this aspect of construct validity of the CAI, this article presents data from two studies on clinical groups of children with speech language impairments and SSDs. The aim of Study 1 is to determine known-group validity by comparing the scores of children with speech language impairments, as diagnosed on the basis of language and intelligence tests, on one task of the CAI (PN) with intelligibility judgments given by SLPs. Study 2 aims to determine the diagnostic power of all four tasks of the CAI by comparing the five CAI factors: PN, NWI-Seg, NWI-Syll, PWV, and MRR (see also Table 4) with a severity judgment of the speech difficulties (mild, moderate, and severe) of children with SSDs.

Study 1

The first study was designed to validate the scores on the PN task of the CAI with intelligibility judgments (good, moderate, poor) in children diagnosed with speech language impairments. For this study, the parameter "percentage of consonants correct" of the PN task is used (PN-PCC), and nonverbal intelligence and language tests are used for the speech language impairment diagnosis.

Method

Ethics, Consent, and Permissions

The research ethics committee of the Radboud University Nijmegen Medical Centre stated that this study does not fall within the remit of the Medical Research Involving Human Subjects Act (Wet medisch-wetenschappelijk onderzoek met mensen; file number: CMO 2016-2985). Therefore, this study can be carried out (in the Netherlands) without an approval by an accredited research ethics committee. Informed consent was obtained from all parents or guardians.

Participants

Ninety-three children aged between 3;0 and 4;0 participated in this study (see Table 1). The sample consisted of 73 boys and 20 girls, representative for the gender distribution in children with speech language impairments. All children attended one of the intervention centers for preschoolers with speech language impairments at the Nederlandse Stichting voor het Dove en Slechthorende Kind, a specialized diagnostic and intervention center for children with hearing loss or speech language impairments. Before admission to the center, these children had been referred to an audiology center (AC) by their family doctor or health care physician on the basis of suspected speech language impairment. At the AC, nonverbal intelligence is assessed by a psychologist, receptive and expressive language tests are administered by an SLP, and hearing status is evaluated by audiometry. Children meet the criteria for referral to a speech language impairment intervention center when they have difficulties in language production and/or language comprehension and/or when their speech is highly unintelligible. Admission takes place if they have a score of at least 1.5 *SD*s below the mean on at least one standardized, norm-referenced language test. Children with hearing loss of 25 dB or more were excluded for this study.

Nonverbal intelligence and language skills were assessed within a period ranging from 3 months before until 3 months after the start of the intervention. If language scores were missing or were older than 3 months at the start of the intervention, language performance was assessed by the SLP of the intervention center within 3 months after the intervention started.

Materials and Procedure

Nonverbal intelligence was assessed with the Snijders-Oomen Nonverbal Intelligence Test 2½-7–Revised (Snijders, Tellegen, Winkel, & Laros, 2003), yielding a nonverbal intelligence quotient (NVIQ). Vocabulary was tested with the Dutch version of the Peabody Picture Vocabulary Test–III (Schlichting, 2005), yielding a vocabulary quotient (QPPVT). The Schlichting Test for Language Comprehension and Language Production (Schlichting & Spelberg, 2010a, 2010b) was used to measure receptive (receptive language quotient: RLQ) and expressive (sentence and word production quotient: SWQ) language skills. These norm-based standard scores or *Q* scores ($M = 100$, $SD = 15$) of each test were used for the analyses.

In addition to the measures for nonverbal intelligence and language, the CAI was administered to all the children (Maassen et al., 2019). For this study, the PN task of the CAI was used. The task was administered by SLPs of the speech language impairment early intervention group, specifically trained in the administration of the CAI. PN contains 60 words, covering the full inventory of vowels, consonants, clusters, and syllable structures of the Dutch language. For this study, the parameter PN-PCC was used for analyses. Individual PN-PCC scores were transformed into *z* scores by subtracting the mean of the normative group and dividing by the standard deviation of the study group; this was done for three age groups (36–39, 40–43, and 44–47 months) separately. The reason for dividing by

Table 1. Number of children per age category and completed tests.

Age category	*N*	Boys	Girls	NVIQ	QPPVT	RLQ	SWQ	PN-PCC-Q
36–39 months	29	23	6	26	25	23	22	29
40–43 months	35	28	7	32	33	21	22	35
44–47 months	29	22	7	26	28	19	17	29
Total	93	73	20	84	86	63	61	93
% Missing values				9.7%	7.5%	32.3%	34.4%	0%

Note. NVIQ = nonverbal intelligence quotient; QPPVT = Peabody Picture Vocabulary Test, vocabulary quotient; RLQ = receptive language quotient; SWQ = sentence and word production quotient; PN-PCC-Q = Computer Articulation Instrument's picture naming percentage consonants correct quotient.

the standard deviation of the study group rather than the standard deviation of the norm group was that the former was approximately three times as large as the latter (18.9 compared to 6.3). Applying the broader confidence intervals of the study group yields the more conservative estimates. z Scores were transformed into Q scores (formula: $Q = 100 + 15*z$) to make them comparable to the cognitive and language scores NVIQ, QPPVT, RLQ, and SWQ.

For each child, the SLP rated the intelligibility on a three-level scale: good, moderate, or poor. The same method is used in the study of Lohmander, Lundeborg, and Persson (2016). Twenty-two children were rated with a “good” intelligibility, 46 were rated as with a “moderate” intelligibility, and 25 children were rated with a “poor” intelligibility.

Statistical Analyses

To test the hypothesis that there is a difference in mean Q scores of the nonverbal intelligence test, language tests, and CAI for the three intelligibility levels, a one-way repeated-measures analysis of variance (ANOVA) was conducted with Q score as a dependent variable, test instrument as a within-subject factor (five levels: NVIQ, QPPVT, RLQ, SWQ, and PN-PCC quotient [PN-PCC-Q]), and intelligibility level as a between-subjects factor (three levels: good, moderate, and poor). Mauchly’s test of sphericity was conducted to test the hypothesis that the variances of differences between conditions are equal. Bonferroni correction was applied for post hoc comparisons. A series of ANOVAs was performed to evaluate differences between Q scores for the three levels of intelligibility. Levene’s test of equality of error variances was conducted to test the homogeneity of variance assumption. Bonferroni correction was applied for post hoc comparisons. Correlations between Q scores and intelligibility levels were calculated with Spearman rank correlation coefficients, and correlations between the Q scores of the different tests were calculated with Pearson rank correlation coefficients. Missing values were replaced by the mean per age group (i.e., mean imputation method). All statistical analyses were performed using SPSS Version 20 for Windows (SPSS Inc.).

Results

Mean Q scores and standard deviations of all tests for the three intelligibility levels are shown in Table 2. Comparing the profiles of Q scores across tests, it was found that, in the levels of moderate and poor intelligibility, on average, children achieved the highest scores on the nonverbal intelligence test, followed by the vocabulary test, the receptive language test, and the expressive language tests. The lowest Q scores were obtained for PN-PCC-Q. In contrast, children with a “good” intelligibility also showed the highest scores for the nonverbal intelligence, but in this group, PN-PCC-Q was higher than the language Q scores, which were approximately equal. Thus, of all Q scores, PN-PCC-Q shows the largest decrease between groups from good to poor intelligibility.

A one-way repeated-measures ANOVA was conducted with the Q scores of the five test instruments as repeated measures and intelligibility level as a between-subjects variable. Mauchly’s test indicated that the assumption of sphericity had been violated, $\chi^2(9) = 58.9$, $p < .001$; therefore, degrees of freedom were corrected using Huynh–Feldt estimates of sphericity ($\varepsilon = .78$). The results show that the within-subject factor “test instrument” was significant, $F(3.10, 278.96) = 79.78$, $p < .001$, effect size or partial $\eta^2 = .47$, which means that the scores on the test instruments were significantly affected by intelligibility level. The between-subjects factor “intelligibility level” was marginally significant, $F(2, 90) = 3.09$, $p = .051$, effect size or partial $\eta^2 = .064$. Post hoc analyses showed that the difference of mean Q scores was not significant between “good” and “moderate” levels ($p = .217$), nor between “moderate” and “poor” levels ($p = .556$), but was significant between “good” and “poor” levels ($M = 6.78$, $SE = 2.47$, $p = .022$). In addition, there was a significant interaction between intelligibility levels and “test instrument,” $F(6.20, 278.96) = 10.00$, $p < .001$, effect size or partial $\eta^2 = .18$. To further examine this interaction, a series of ANOVAs was conducted to test the differences between the three intelligibility levels for the Q scores of each test instrument separately. There was no significant difference between intelligibility levels for NVIQ, $F(2, 90) = 0.47$, $p = .626$; QPPVT, $F(2, 90) = 0.87$, $p = .421$; RLQ, $F(2, 90) = 0.43$, $p = .650$; or SWQ, $F(2, 90) = 3.07$, $p = .051$. For the latter, marginally significant factor SWQ, post hoc analyses revealed a significant mean difference between “good” and “poor” levels ($p = .047$) and no significant mean differences between “good” and “moderate” levels ($p = .276$) or “moderate” and “poor” levels ($p = .795$). For PN-PCC-Q, the Levene’s test for equality of variances was significant, indicating that the requirement of homogeneity of variance was violated. Therefore, the Welch F ratio was calculated, showing that the difference in mean PN-PCC-Q between intelligibility levels was significant, $F(2, 51.28) = 69.48$, $p \leq .001$.

Table 3 shows correlations between intelligibility and Q scores. A strong, significant correlation was found between PN-PCC-Q and intelligibility (Spearman $r(93) = .69$, $p < .001$), which is in the expected direction: PN-PCC-Q decreases when the intelligibility level decreases. No other Q scores, not even the expressive language score SWQ, correlated significantly with intelligibility or with PN-PCC-Q. There were weak, significant correlations between the outcome of the nonverbal intelligence test and language tests and moderate correlations among the language tests, with correlations between RLQ and SWQ and between QPPVT and RLQ being moderate and the correlation between QPPVT and SWQ being weak. No significant correlations were found between PN-PCC-Q and the Q scores of the nonverbal intelligence test and language tests. Inspection of the scatter plots did not reveal any outliers.

Study 2

The second study aims to determine the diagnostic power of all four tasks of the CAI. For this, the relation

Table 2. Mean *Q* scores for the nonverbal intelligence, language, and speech tests.

Intelligibility score	*N*	NVIQ		QPPVT		RLQ		SWQ		PN-PCC-Q	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Good	22	102.8	11.6	84.7	18.8	78.6	11.3	78.2	10.4	92.5	5.99
Moderate	46	99.9	11.2	89.9	16.6	80.9	13.7	74.0	9.50	73.4	11.3
Poor	25	100.9	11.4	90.9	18.2	82.2	14.0	71.3	9.33	62.5	14.1

Note. NVIQ = nonverbal intelligence quotient; QPPVT = Peabody Picture Vocabulary Test, vocabulary quotient; RLQ = receptive language quotient; SWQ = sentence and word production quotient; PN-PCC-Q = Computer Articulation Instrument's picture naming percentage consonants correct quotient.

between the five CAI factors (PN, NWI-Seg, NWI-Syll, word and nonword PWV, and MRR) and clinical judgments of severity of the speech disorder by the SLPs is investigated.

Method

Ethics, Consent, and Permissions

The ethics approval for Study 1 also applied to Study 2.

Participants

The participants in Study 2 were 41 children with an age range from 3;0 to 6;4, with 26 boys and 15 girls. For this study, children with SSDs were recruited from several institutions: 19 children from primary health care services, one child from an AC, and 21 children from a special school for children with language and hearing impairments. All parents or caregivers were given an information letter. After obtaining the signed parental consent form, the child was included in the study.

The parents or caregivers of all 41 children were asked to provide information about the children's hearing status. They were asked whether the child had a history of hearing problems, if hearing problems had been recorded during the regular governmental (neonatal) hearing screening, and, if available, if they could provide us with hearing acuity data (pure-tone thresholds). Thirty children passed a bilateral hearing screening at 20 dB. Parents or caregivers of the other 11 children reported no history of hearing problems and no hearing problems recorded during the regular governmental (neonatal) hearing screening.

Prior to the procedures of this study, a speech diagnosis was reported by the SLP of the child, based on clinical observation and a standard speech-language protocol, including standardized language tests. Speech was observed with different instruments. Until now, for the Dutch language, no standardized and normalized speech assessment is available. All children were diagnosed with SSDs, most of them (n = 36) with a PD, two children with CAS, and three children with an unknown diagnosis because no details were available about the children's speech apart from the fact that their SSD was severe. Differential diagnosis was part of the clinical reasoning process of the SLP and was done based on diagnostic criteria described in studies such as Forrest (2003) and Shriberg and Kwiatkowski (1994).

Materials and Procedure

For this study, all participants were tested on their speech skills with the CAI. All four tasks (PN, NWI, WR and NWR, and MRR) were administered. Both the administration of the tests and the analyses of the speech are computer implemented. Table 4 shows the parameters used to assess task performance; a detailed description of the CAI and these parameters, as well as a description of the normative data set, is presented in Maassen et al. (2019) and van Haaften et al. (2019); for all parameters, percentile

Table 3. Spearman and Pearson rank correlations between intelligibility levels and *Q* scores and between *Q* scores (*N* = 93).

Intelligibility level and Q scores		Intelligibility level	NVIQ	QPPVT	RLQ	SWQ	PN-PCC-Q
Intelligibility level	Spearman *r*	1	.027	−.14	−.11	.20	.69**
NVIQ	Pearson *r*	—	1	.36**	.31**	.35**	.10
QPPVT	Pearson *r*		—	1	.52**	.36**	−.22
RLQ	Pearson *r*			—	1	.48**	−.15
SWQ	Pearson *r*				—	1	.21
PN-PCC-Q	Pearson *r*					—	1

Note. NVIQ = nonverbal intelligence quotient; QPPVT = Peabody Picture Vocabulary Test, vocabulary quotient; RLQ = receptive language quotient; SWQ = sentence and word production quotient; PN-PCC-Q = Computer Articulation Instrument's picture naming percentage consonants correct quotient.

*Correlation of factor scores is significant at the .05 level (two-tailed). **Correlation of factor scores is significant at the .01 level (two-tailed).

Table 4. Computer Articulation Instrument parameters per speech task and extracted factors.

Task	Factor		Parameter	
PN	PN		PCCI	Percentage of consonants correct in syllable-initial position
			PVC	Percentage of vowels correct
			Level 5	Percentage of correct consonants /l/ and /R/
			RedClus	Percentage of reduction of initial consonant clusters from two consonants to one
			CCVC	Percentage of correct syllable structure CCVC (C = consonant, V = vowel)
NWI	NWI-Seg		PCCI	Percentage of consonants correct in syllable-initial position
			PVC	Percentage of vowels correct
			Level 4	Percentage of correct consonants /b/, /f/, and /ʋ/
			Level 5	Percentage of correct consonants /l/ and /R/
			CVC	Percentage of correct syllable structure CVC
	NWI-Syll		RedClus	Percentage of reduction of initial consonant clusters from two consonants to one
			CCVC	Percentage of correct syllable structure CCVC
WR	PWV		PWV Word	Proportion of whole-word variability: word repetition
NWR			PWV Nonword	Proportion of whole-word variability: nonword repetition
MRR		MRR-Mono	MRR-pa	Number of syllables per second of sequence /pa/
			MRR-ta	Number of syllables per second of sequence /ta/
			MRR-ka	Number of syllables per second of sequence /ka/
		MRR-BiTri	MRR-pataka	Number of syllables per second of sequence /pataka/
			MRR-pata	Number of syllables per second of sequence /pata/
			MRR-taka	Number of syllables per second of sequence /taka/

Note. PN = picture naming; NWI = nonword imitation; WR = word repetition; NWR = nonword repetition; MRR = maximum repetition rate; PN = factor score of all parameters of picture naming; NWI-Seg = factor score of the segmental parameters of nonword imitation; NWI-Syll = factor score of the syllable structure parameters of nonword imitation; PWV = factor score of the two PWV parameters of word and nonword repetition; MRR-Mono = factor score of the monosyllabic items of maximum repetition rate parameters; MRR-BiTri = factor score of the bisyllabic and trisyllabic items of maximum repetition rate parameters.

scores can be determined. A factor analysis on all 20 parameters of the normative data, obtained from a total number of 1,524 children, yielded five factors: (a) PN, (b) NWI-Seg, (c) NWI-Syll, (d) PWV of words and nonwords, and (e) MRR (van Haaften et al., 2019). For this study, factor scores were calculated based on the factor weights obtained from this factor analysis. Because there were many missing values in the MRR task (see below), separate factor scores were calculated on only the monosyllabic MRR sequences (/papa../, /tata../, /kaka../; yielding factor MRR-Mono) and the bisyllabic (/pata../, /taka../) and trisyllabic (/pataka../) sequences, yielding factor MRR-BiTri.

Prior to the administration of the CAI, severity of the SSDs was judged by the child's SLP (N = 11) on a severity scale with three categories—mild, moderate, and severe—following the categories proposed by Dodd (1995c). An SLP rated the severity of an SSD as *mild* when a child is mostly intelligible in spontaneous speech but errors are obvious and distracting from content. The severity was rated *moderate* when single words are often intelligible in context but connected speech is often difficult to understand, particularly out of context. The category *severe* was rated when most utterances are unintelligible on the first meeting. Also, the persistence of the speech disorder and the consequences on communication abilities were taken into account when rating severity. The category "moderate" was scored for 14 children, and 27 children were scaled as "severe." None of the children was scaled as having a "mild" speech disorder. Therefore, the statistical analyses of this study are based on two severity categories: moderate and severe. Table 5 shows the distribution of the participants in the three severity categories by speech diagnosis.

The tasks of the CAI were administered by (candidate) SLPs specifically trained in the administration of the CAI.

Statistical Analyses

The factor PWV had two missing values, and these were replaced by the overall PWV mean (M = −1.20; i.e., mean imputation method). Much more missing data were observed for the MRR tasks, due to speech–motor difficulties and/or shyness or inattentiveness of the child; also, a few recordings could not be analyzed due to the low acoustic quality. Of the total number of 41 children, only 23 produced at least two monosyllabic sequences correctly (44% missing),

Table 5. Speech diagnosis by severity categories.

Severity category	Speech disorder			
	PD	CAS	Unknown	Total
Mild	0	0	0	0
Moderate	13	1	0	14
Severe	23	1	3	27
Total	36	2	3	41

Note. PD = phonological disorder; CAS = childhood apraxia of speech.

and only nine of these 23 (amounting to 78% missing data) produced at least two of the bisyllabic or trisyllabic sequences. Because of this large number of missing values, no imputation was applied, but a separate analysis was conducted instead on the group of 23 children. The 14 children who were not able to produce the bisyllabic or trisyllabic sequences were assigned the lowest z score, such that failure to produce these sequences was marked as poor performance. One-way repeated-measures ANOVAs were conducted to test the hypothesis that there is a difference in CAI factors for the two severity categories, comprising two levels: "moderate" and "severe." Because of the missing data in factors MRR-Mono and MRR-BiTri, the first analysis was conducted on the four remaining factors: PN, NWI-Seg, NWI-Syll, and PWV. Subsequently, a one-way repeated-measures ANOVA was conducted with six CAI factors, including MRR-Mono and MRR-BiTri. Mauchly's tests of sphericity were conducted to test the hypothesis that the variances of differences between conditions are equal. Next, if in the ANOVA either severity level or the interaction between severity level and CAI factor was significant, a series of independent t tests was conducted to evaluate the difference in factor scores between the moderate and severe groups for each of the four or six CAI factors separately. Levene's test of equality of error variances was conducted to test the homogeneity of variance assumption. Correlations between CAI factors and severity categories were calculated by Spearman rank correlation coefficients (r), and correlations between the CAI factors were assessed by calculating Pearson rank correlation coefficients (r). All statistical analyses were performed using SPSS Version 20 for Windows (SPSS Inc.).

Results

Table 6 shows that, on average, children with a speech disorder of moderate severity have higher factor scores on PN, NWI-Seg, NWI-Syll, and PWV, than children with a severe speech disorder. For the children with a severe speech disorder, mean factor scores ranged from −1.13 to −1.72; and for the children with moderate severity, between −0.18 and −1.07. Thus, all mean scores were below the population average.

First, a one-way repeated-measures ANOVA with the four CAI factors PN, NWI-Seg, NWI-Syll, and PWV was conducted. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(5) = 15.13$, $p = .010$; therefore, degrees of freedom were corrected using Huynh–Feldt estimates of sphericity ($\varepsilon = .91$). The results show that the within-subject factor "CAI factors" was significant, $F(2.74, 106.96) = 18.29$, $p < .001$, effect size or partial $\eta^2 = .32$, indicating that the factor scores of the CAI were significantly affected by the severity of the speech disorder. The between-subjects factor "severity category" was also significant, $F(1, 39) = 11.98$, $p = .001$, effect size or partial $\eta^2 = .24$; there was a significant difference in factor scores between the children with moderate and severe speech disorders. There was also a significant interaction between CAI factors and severity categories, $F(2.74, 106.96) = 3.70$, $p = .017$, effect size or partial $\eta^2 = .087$. To further examine this interaction, a series of independent t tests was conducted to test the differences between the two severity categories for each CAI factor separately. Significantly lower factor scores for the severe versus moderate groups were found for PN, $t(39) = 3.62$, $p = .001$; NWI-Seg, $t(39) = 3.21$, $p = .003$; and NWI-Syll, $t(39) = 3.67$, $p = .001$. No significant difference was found between the mean factor scores of the moderate and severe groups for the CAI factor PWV, $t(39) = 1.11$, $p = .27$.

The second one-way repeated-measures ANOVA was conducted with all six CAI factors, including MRR-Mono and MRR-BiTri, on 23 children with complete data on these factors (see Table 7). A one-way repeated-measures ANOVA was conducted with these CAI factors: PN, NWI-Seg, NWI-Syll, PWV, MRR-Mono, and MRR-BiTri. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(14) = 32.99$, $p = .003$; therefore, degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon = .87$). Like the analysis with four factors, the results show that the six factor scores of the CAI were significantly affected by the severity level of the speech disorder; the within-subject factor "CAI factors" was significant,

Table 6. Means and standard deviations of the factor scores of four Computer Articulation Instrument factors per severity category.

Severity category		PN	NWI-Seg	NWI-Syll	PWV
Moderate	*N*	14	14	14	14
	M	−1.07	−0.88	−0.18	−1.03
	SD	0.52	0.83	0.75	0.81
Severe	*N*	27	27	27	27
	M	−1.72	−1.69	−1.13	−1.29
	SD	0.56	0.73	0.80	0.67
Total	*N*	41	41	41	41
	M	−1.45	−1.42	−0.81	−1.20
	SD	0.62	0.85	0.89	0.72

Note. PN = factor score of all parameters of picture naming; NWI-Seg = factor score of the segmental parameters of nonword imitation; NWI-Syll = factor score of the syllable structure parameters of nonword imitation; PWV = factor score of the two PWV parameters of word and nonword repetition.

Table 7. Means and standard deviations of the factor scores of six Computer Articulation Instrument factors per severity category.

Severity category		PN	NWI-Seg	NWI-Syll	PWV	MRR-Mono	MRR-BiTri
	N	10	10	10	10	10	10
Moderate	*M*	–1.05	–0.80	–0.29	–0.97	–1.12	–1.15
	SD	0.46	0.69	0.80	0.85	0.89	1.56
	N	13	13	13	12	13	13
Severe	*M*	–1.52	–1.45	–1.14	–1.28	–0.60	–2.31
	SD	0.67	0.84	0.75	0.75	0.85	0.06
	N	23	23	23	22	23	23
Total	*M*	–1.31	–1.17	–0.77	–1.14	–0.82	–1.81
	SD	0.62	0.83	0.87	0.80	0.89	1.16

Note. PN = factor score of all parameters of picture naming; NWI-Seg = factor score of the segmental parameters of nonword imitation; NWI-Syll = factor score of the syllable structure parameters of nonword imitation; PWV = factor score of the two PWV parameters of word and nonword repetition; MRR-Mono = factor score of the monosyllabic items of maximum repetition rate parameters; MRR-BiTri = factor score of the bisyllabic and trisyllabic items of maximum repetition rate parameters.

$F(4.3, 90.9) = 6.40$, $p < .001$, effect size or partial $\eta^2 = .23$. The between-subjects factor "severity category" was also significant, $F(1, 21) = 4.60$, $p = .04$, effect size or partial $\eta^2 = .18$, as well as the interaction between CAI factors and severity categories, $F(4.3, 90.9) = 4.17$, $p = .003$, effect size or partial $\eta^2 = .17$. To further examine this interaction, independent *t* tests were conducted to test the differences between the two severity categories for all six factors. For NWI-Syll, $t(21) = 2.61$, $p = .016$, and MRR-BiTri, $t(0.0) = 2.35$, $p = .043$, the differences between the mean factor scores of the moderate and severe groups reached significance. No significant difference was found between the severity groups for PWV. For PN and NWI-Seg, the differences were only marginally significant in this second analysis, most likely due to less power as compared to the first analysis. It is remarkable that there is no difference between the moderate and severe groups for MRR-Mono, but there is a large significant difference for MRR-BiTri. We will come back to this issue in the general discussion.

Table 8 shows correlations between severity category and CAI factors. Moderate, significant correlations were found between severity category and PN, NWI-Seg, and NWI-Syll. Children with a severe disorder had lower CAI factor scores. The factor scores of PN, NWI-Seg, and NWI-Syll showed strong correlations; the correlations with PWV and MRR-BiTri were weak to moderate. No significant correlations were found between MRR-Mono and any other CAI factor.

Discussion

The CAI is a computer-based assessment for speech production with a range of speech tasks that reflect different levels of processing (phonological and speech motor skills), and it provides normative data based on a sample of 1,524 children in the age range of 2;0–6;11. A previous study on psychometric characteristics of the CAI revealed sufficient interrater reliability, test–retest reliability, and construct validity (van Haaften et al., 2019). In this current article, we report known-group validity, based on the outcome of two studies in children with speech language impairment and SSDs.

Table 8. Spearman rank correlations and Pearson correlations between severity category and Computer Articulation Instrument factors and between Computer Articulation Instrument factors.

Severity category and CAI factors		Severity category	PN	NWI-Seg	NWI-Syll	PWV	MRR-Mono	MRR-BiTri
	N	41	41	41	41	41	23	23
Severity category	Spearman *r*	1	–.53**	–.50**	–.50**	–.19	.28	–.32
PN	Pearson *r*	—	1	.80**	.81**	.39*	–.09	.41*
NWI-Seg	Pearson *r*		—	1	.68**	.60**	.12	.53*
NWI-Syll	Pearson *r*			—	1	.51**	–.03	.44*
PWV	Pearson *r*				—	1	–.07	.21
MRR-Mono	Pearson *r*					—	1	–.02
MRR-BiTri	Pearson *r*						—	1

Note. CAI = computer articulation instrument; PN = factor score of all parameters of picture naming; NWI-Seg = factor score of the segmental parameters of nonword imitation; NWI-Syll = factor score of the syllable structure parameters of nonword imitation; PWV = factor score of the two PWV parameters of word and nonword repetition; MRR-Mono = factor score of the monosyllabic items of maximum repetition rate parameters; MRR-BiTri = factor score of the bisyllabic and trisyllabic items of maximum repetition rate parameters.

*Correlation of factor scores is significant at the .05 level (two-tailed). **Correlation of factor scores is significant at the .01 level (two-tailed).

The known-group validity of the CAI was supported by the results of Study 1. These results confirm the hypothesis that PN-PCC-Q is significantly affected by intelligibility level. There was a significant difference between the intelligibility levels with respect to the PCC parameter of the PN task of the CAI, and there was a highly significant correlation between the intelligibility levels and PN-PCC-Q in the expected direction. Correlations between PCC and intelligibility measures were also found in previous studies (Lagerberg et al., 2015; McLeod, Harrison, & McCormack, 2012; Neumann, Rietz, & Stenneken, 2017). In the study of McLeod et al. (2012), significant correlations were found between PCC (measured with the Phonology subtest of the Diagnostic Evaluation of Articulation and Phonology) and the outcome of the Intelligibility in Context Scale. Unfortunately, the Intelligibility in Context Scale could not be administered in our study, because the children in Study 1 fell out of its age range (too young). Therefore, the intelligibility was scored by the SLPs on a scale with three levels: good, moderate, and poor. In Study 1 and Study 2, subjective judgments of SLPs with ordinal scales were used. Due to this subjectivity, no optimal objective measurements were collected, which is a limitation of this study. No reliability measures are reported for these scales. However, it is a common way to judge children's speech, and they are used in several other studies (Gordon-Brannan & Hodson, 2000; Lohmander et al., 2016). Further validation studies are needed to corroborate the diagnostic value of the CAI. This study with "expert judgment" is the first step in this validation process. Different studies describe that experienced listeners tend to give higher intelligibility ratings than inexperienced listeners (Doyle, Swift, & Haaf, 1989; Landa et al., 2014). In the current study, the ratings were assigned by SLPs who are experienced listeners. As a consequence, the rating "poor intelligibility" must be considered as an indication of a serious speech difficulty. It emphasizes the validity of the strongly related parameter PN-PCC-Q. The results of our study showed a quite stable pattern of nonverbal intelligence and language scores in the children with a speech language impairment across intelligibility levels. Intelligibility level shows no or only a very weak, nonsignificant correlation with the outcomes on the nonverbal intelligence and language tests; similarly, no or a very weak, nonsignificant correlation was found between PN-PCC-Q and the outcomes on the nonverbal intelligence and language tests. The results of these correlations show that the PCC of PN of the CAI measures a distinct aspect of the language domain. This corresponds to the subtypes described by Van Weerdenburg et al. (2006), in which children with an SSD are one of the four distinct subtypes.

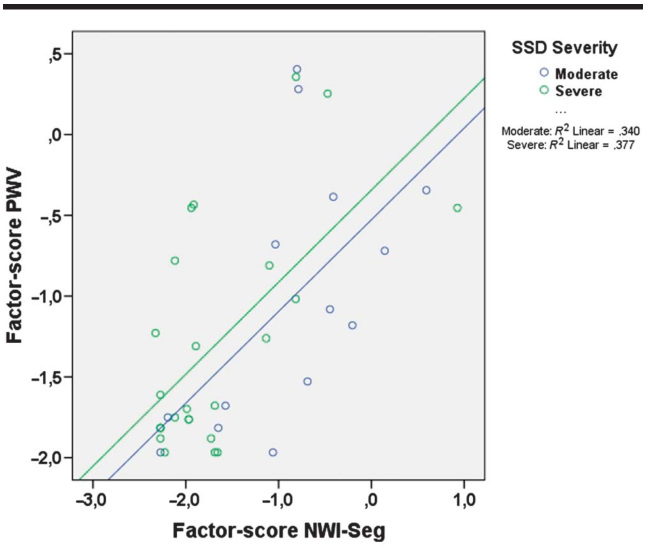
Study 2 supports the diagnostic power of the CAI factors in a group of children with SSDs. All children, with either a moderate or severe SSD, showed scores below average on the CAI factors PN, NWI-Seg and NWI-Syll, PWV, MRR-Mono, and MRR-BiTri, with mean factor scores being between −0.77 and −1.81.

Comparison of four CAI factors (without MRR) revealed significant differences among these factors and between the two severity categories. The severity of the speech disorder is mainly expressed in the parameters of PN and NWI, as shown by the significant difference between the moderate and severe groups for the CAI factors PN, NWI-Seg, and NWI-Syll, whereas PWV is stable across the two groups. These results suggest that especially PN and NWI are the most sensitive tasks to diagnose SSDs. This is in line with other authors who stated that NWI, in which articulatory competence is tested separately from lexical knowledge, is an important part of an assessment battery for children with SSDs (Vance, Stackhouse, & Wells, 2005). Other authors have also suggested to not only use PN in a speech assessment but also include an NWI task to gain better insight in the speech production of a child (Geronikou & Rees, 2016; Hodges, Baker, Munro, & McGregor, 2017). NWI is also associated with phonological short-term memory (Gathercole, 2006). Poor performance on NWI can be influenced by difficulties with phonological short-term memory and not just speech production difficulties. Krishnan et al. (2017) suggest that NWI skills have a unique role in the process of remembering and reproducing novel words. They found that NWI abilities were associated with oromotor praxis, reading fluency, and audiovisual sequence reproduction accuracy. The finding that PWV is relatively stable across severity groups might be related to the multiple origins of inconsistency. As elaborated in the introduction, inconsistency could indicate unstable lexical representations, an unstable phonological system, or unstable motor planning as is typical for CAS.

When all six CAI factors were compared (including MRR), significant differences were found among the six factors and the two severity categories. Differences between the moderate and severe groups were found for PN, NWI-Seg, NWI-Syll, and MRR-BiTri. Remarkably, no difference between the moderate and severe groups was found for MRR-Mono, whereas there was a significant difference between the moderate and severe groups for MRR-BiTri. The severe group showed the lowest z score for MRR-BiTri (−2.31) when compared with the other CAI factors. These results imply that MRR-BiTri is an important factor in diagnosing SSDs, such as PN and NWI. MRR-BiTri is especially useful in differential diagnosis of SSDs with a motor origin (CAS and dysarthria), as mentioned in other studies (Rvachew, Hodge, & Ohberg, 2005; Thoonen, Maassen, Wit, Gabreëls, & Schreuder, 1996). The fact that PN, NWI, and MRR-BiTri of the CAI were the most affected in the severe speech disorder group underlines the importance of these tasks in diagnosing SSDs. No differences between the two severity groups were found for the factors PWV and MRR-Mono. They correlate less with the SLPs' judgments of severity than the other factors. Nevertheless, the mean factor scores are below average in the SSD groups as compared to typically developing children with the same age. This indicates that these tasks do contribute to the diagnostic differentiation between typical and atypical development. In studies on speech development, speech variability, as assessed with the WR and NWR tasks, has been found to be relatively high in young typically developing children (2- and 3-year-olds; Sosa, 2015), and such variability decreases with age (Holm, Crosble, & Dodd, 2007). In a previous study (van Haaften et al., 2019), we also found minor decreases of the PWV with age. Increased variability has also been associated with certain types of speech disorders, such as CAS (Davis et al., 1998; Dodd, 1995b; Forrest, 2003; Holm et al., 2007; Iuzzini-Seigel et al., 2017) and inconsistent PDs (Dodd, 1995b). In this study, PWV shows a mean below-average factor score and a moderate to strong correlation (.39–.60) to the PN and NWI factors, although the PWV scores for moderate and severe disorders do not differ. To get a better understanding of these complex relations, a scatter plot of PWV and NWI-Seg factor scores was made (see Figure 2). Regression lines show a small difference in PWV between moderate and severe disorders; interestingly, for both severity groups, the correlation with NWI-Seg is equally strong. This suggests that PWV can serve as a diagnostic marker for SSDs; validation studies with other speech and language diagnoses need to be conducted.

MRR performance of monosyllabic sequences shows no relation with the other task parameters, suggesting that MRR-Mono assesses an independent aspect of speech production. This is in accordance with such studies as the one by Staiger, Schölderle, Brendel, Bötzel, and Ziegler (2017), who concluded, from factor analyses of speech data from patients with neurological movement disorders as compared to control subjects, that speech tasks and oral motor tasks such as rapid syllable repetition measure separate traits. Krishnan et al. (2017) studied the correlation between NWI and other tasks. They also found no correlation between MRR-Mono and NWI, whereas an alternate MRR task (such as MRR-BiTri) correlated significantly with NWI. From the perspective of a process-oriented approach, Maassen and Terband (2015) argued that MRR, being a pure motor task that does not require any knowledge of words, syllables, or phonemes, can be used to assess speech motor skills. Still, like PWV, mean MRR-Mono factor scores are below the population average and thus, like PWV, might serve as a diagnostic marker for SSDs. However, in contrast to PWV, MRR-Mono does not correlate with severity. Further studies are needed to delineate the role of the purely repetitive (MRR-Mono) and sequential (MRR-BiTri) variants in SSDs.

Figure 2. Scatter plot of the segmental quality of nonword imitation (NWI-Seg) and word and nonword proportion of whole-word variability (PWV factor scores), showing the correlations for both groups of children with moderate and severe speech sound disorders (SSDs). Although the difference in PWV between the two groups is small, the correlations with NWI-Seg are moderate to strong.



This study yields strong indications that comparison of the performance on the different speech tasks of the CAI provides information on the underlying speech processing difficulties of children with SSDs. Interestingly, the children with SSDs show a distinct factor structure, which differs from that of the normative study. As mentioned in the introduction, in the normative study on 1,524 typically developing children, weak and very weak correlations between factor scores were found, from which it can be concluded that the CAI factors represent independent components of the speech production process. Aligned with psycholinguistic models, such as Levelt's model, the current study describes the speech profile of a group of children with SSDs by conducting different speech tasks covering all different speech processes (phonological and speech motor skills). A limitation of this study is the use of a heterogeneous group of children with SSDs, without analyzing the results of different subgroups. This is an important next step in process-oriented diagnostics. The crucial statistical remark to be made here is that factor analysis is based not on average skills but on variability in skills and especially covariance. It can be argued that, in a typical population, variability in skills is not caused by specific underlying factors but rather reflects random noise. In contrast, in an atypical population such as children with SSDs, underlying deficits can cause large covariance if task requirements show overlap; analyzing this structure of overlapping and nonoverlapping task performances is the first step in process-oriented diagnostics. Future investigations are needed to compare subgroups of children with different types of SSDs, such that more profiles of CAI factors can be determined to further reveal the proximal causes of SSDs.

Following the results of the study, the most important implication for clinical practice is to distinguish typical speech development from atypical speech development by the administration of different speech tasks, such as incorporated in the CAI. This allows for process-oriented diagnostics, which is important for targeted intervention in children with SSDs.

Acknowledgments

The authors did not receive any grant funding for this study, but this research was partially supported by Het Revalidatiefonds (The Fund for Rehabilitation). The authors wish to thank all the children who participated, their parents, and all the SLPs for their valuable help with the administration and scoring of the tests.

References

Baker, E., Croot, K., McLeod, S., & Paul, R. (2001). Psycholinguistic models of speech development and their application to clinical practice. *Journal of Speech, Language, and Hearing Research, 44*(3), 685–702. https://doi.org/10.1044/1092-4388(2001/055)

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., Adams, C., Archibald, L., . . . House, A. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *The Journal of Child Psychology and Psychiatry, 58*(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

Davis, B. L., Jakielski, K. J., & Marquardt, T. P. (1998). Developmental apraxia of speech: Determiners of differential diagnosis. *Clinical Linguistics & Phonetics, 12*(1), 25–45. https://doi.org/10.3109/02699209808985211

Dodd, B. (1995a). Children's acquisition of phonology. In B. Dodd (Ed.), *Differential diagnosis and treatment of children with speech disorder* (pp. 21–48). London, England: Whurr.

Dodd, B. (1995b). Children with speech disorder: Defining the problem. In B. Dodd (Ed.), *Differential diagnosis and treatment of children with speech disorder*. (pp. 1–19). London, England: Whurr.

Dodd, B. (1995c). Procedures for classification of subgroups of speech disorder. In B. Dodd (Ed.), *Differential diagnosis and treatment of children with speech disorder* (pp. 49–64). London, England: Whurr.

Dodd, B. (2011). Differentiating speech delay from disorder: Does it matter? *Topics in Language Disorders, 31*(2), 96–111. https://doi.org/10.1097/TLD.0b013e318217b66a

Dodd, B. (2014). Differential diagnosis of pediatric speech sound disorder. *Current Developmental Disorders Reports, 1*(3), 189–196. https://doi.org/10.1007/s40474-014-0017-3

Doyle, P. C., Swift, E. R., & Haaf, R. G. (1989). Effects of listener sophistication on judgments of tracheoesophageal talker intelligibility. *Journal of Communication Disorders, 22*(2), 105–113. https://doi.org/10.1016/0021-9924(89)90027-0

Eadie, P., Morgan, A., Ukoumunne, O. C., Ttofari Eecen, K., Wake, M., & Reilly, S. (2015). Speech sound disorder at 4 years: Prevalence, comorbidities, and predictors in a community cohort of children. *Developmental Medicine & Child Neurology, 57*(6), 578–584. https://doi.org/10.1111/dmcn.12635

Forrest, K. (2003). Diagnostic criteria of developmental apraxia of speech used by clinical speech-language pathologists. *American Journal of Speech-Language Pathology, 12*(3), 376–380. https://doi.org/10.1044/1058-0360(2003/083)

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 27*(4), 513.

Geronikou, E., & Rees, R. (2016). Psycholinguistic profiling reveals underlying impairments for Greek children with speech disorders. *Child Language Teaching and Therapy, 32*(1), 95–110. https://doi.org/10.1177/0265659015583915

Gordon-Brannan, M., & Hodson, B. (2000). Intelligibility/severity measurements of prekindergarten children's speech. *American Journal of Speech-Language Pathology, 9*(2), 141–150. https://doi.org/10.1044/1058-0360.0902.141

Hodges, R., Baker, E., Munro, N., & McGregor, K. K. (2017). Responses made by late talkers and typically developing toddlers during speech assessments. *International Journal of Speech-Language Pathology, 19*(6), 587–600. https://doi.org/10.1080/17549507.2016.1221452

Holm, A., Crosble, S., & Dodd, B. (2007). Differentiating normal variability from inconsistency in children's speech: Normative data. *International Journal of Language & Communication Disorders, 42*(4), 467–486. https://doi.org/10.1080/13682820600988967

Iuzzini-Seigel, J., Hogan, T. P., & Green, J. R. (2017). Speech inconsistency in children with childhood apraxia of speech, language impairment, and speech delay: Depends on the stimuli. *Journal of Speech, Language, and Hearing Research, 60*(5), 1194–1210. https://doi.org/10.1044/2016_JSLHR-S-15-0184

Krishnan, S., Alcock, K. J., Carey, D., Bergström, L., Karmiloff-Smith, A., & Dick, F. (2017). Fractionating nonword repetition: The contributions of short-term memory and oromotor praxis are different. *PLOS ONE, 12*(7), 1–18. https://doi.org/10.1371/journal.pone.0178356

Lagerberg, T. B., Hartelius, L., Johnels, J. Å., Ahlman, A.-K., Börjesson, A., & Persson, C. (2015). Swedish Test of Intelligibility for Children (STI-CH)—Validity and reliability of a computer-mediated single word intelligibility test for children. *Clinical Linguistics & Phonetics, 29*(3), 201–215. https://doi.org/10.3109/02699206.2014.987925

Landa, S., Pennington, L., Miller, N., Robson, S., Thompson, V., & Steen, N. (2014). Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. *International Journal of Speech-Language Pathology, 16*(4), 408–416. https://doi.org/10.3109/17549507.2014.927922

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (2000). Prevalence and natural history of primary speech and language delay: Findings from a systematic review of the literature. *International Journal of Language & Communication Disorders, 35*(2), 165–188.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1–75.

Lohmander, A., Lundeborg, I., & Persson, C. (2016). SVANTE—The Swedish Articulation and Nasality Test—Normative data and a minimum standard set for cross-linguistic comparison. *Clinical Linguistics & Phonetics, 31,* 137–154. https://doi.org/10.1080/02699206.2016.1205666

Maassen, B., & Terband, H. (2015). Process-oriented diagnosis of childhood and adult apraxia of speech (CAS and AOS). In M. A., Redford (Ed.), *The handbook of speech production* (1st ed., pp. 331–350). United Kingdom: Wiley.

Maassen, B., van Haaften, L., Diepeveen, S., Terband, H., van den Engel-Hoek, L., Veenker, T., & De Swart, B. (2019). *Computer articulation instrument*. Amsterdam, the Netherlands: Boom Uitgevers.

Macrae, T., Tyler, A. A., & Lewis, K. E. (2014). Lexical and phonological variability in preschool children with speech sound disorder. *American Journal of Speech-Language Pathology, 23,* 27–35. https://doi.org/10.1044/1058-0360(2013/12-0037)

McLeod, S., Harrison, L. J., & McCormack, J. (2012). The intelligibility in context scale: Validity and reliability of a subjective rating measure. *Journal of Speech, Language, and Hearing Research, 55*(2), 648–656. https://doi.org/10.1044/1092-4388(2011/10-0130)

Mullen, R., & Schooling, T. (2010). The national outcomes measurement system for pediatric speech-language pathology. *Language, Speech, and Hearing Services in Schools, 41*(1), 44–60. https://doi.org/10.1044/0161-1461(2009/08-0051)

Neumann, S., Rietz, C., & Stenneken, P. (2017). The German Intelligibility in Context Scale (ICS-G): Reliability and validity evidence. *International Journal of Language & Communication Disorders, 52*(5), 585–594. https://doi.org/10.1111/1460-6984.12303

Nijland, L. (2003). *Developmental apraxia of speech: Deficits in phonetic planning and motor programming*. Wageningen, the Netherlands: Ponson & Looijen.

Nijland, L., Maassen, B., & van der Meulen, S. (2003). Evidence of motor programming deficits in children diagnosed with DAS. *Journal of Speech, Language, and Hearing Research, 46*(2), 437–450. https://doi.org/10.1044/1092-4388(2003/036)

Nijland, L., Maassen, B., van der Meulen, S., Gabreëls, F., Kraaimaat, F. W., & Schreuder, R. (2003). Planning of syllables in children with developmental apraxia of speech. *Clinical Linguistics & Phonetics, 17*(1), 1–24. https://doi.org/10.1080/0269920021000050662

Nijland, L., Terband, H., & Maassen, B. (2015). Cognitive functions in childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research, 58*(3), 550–565. https://doi.org/10.1044/2015_JSLHR-S-14-0084

Portney, L., & Watkins, M. (2009). *Foundations of clinical research: Applications to practice*. Upper Saddle River, NJ: Prentice Hall.

Rvachew, S., Hodge, M., & Ohberg, A. (2005). Obtaining and interpreting maximum performance tasks from children: A tutorial. *Journal of Speech-Language Pathology and Audiology, 29*(4), 146–157.

Schlichting, L. (2005). *Peabody Picture Vocabulary Test–III Dutch version (PPVT-III-NL)*. Amsterdam, the Netherlands: Pearson Assessment and Information B.V.

Schlichting, L., & Spelberg, H. L. (2010a). *Schlichting Test voor Taalbegrip* [Schlichting Test for Language Comprehension]. Houten, the Netherlands: Bohn Stafleu van Loghum.

Schlichting, L., & Spelberg, H. L. (2010b). *Schlichting Test voor Taalproductie-II* [Schlichting Test for Language Production]. Houten, the Netherlands: Bohn Stafleu van Loghum.

Shriberg, L. D., & Kwiatkowski, J. (1994). Developmental phonological disorders I: A clinical profile. *Journal of Speech and Hearing Research, 37*(5), 1100–1126.

Shriberg, L. D., Strand, E. A., Fourakis, M., Jakielski, K. J., Hall, S. D., Karlsson, H. B., . . . Wilson, D. L. (2017). A diagnostic marker to discriminate childhood apraxia of speech from speech delay: I. Development and description of the pause marker. *Journal of Speech, Language, and Hearing Research, 60*(4), 1096–1117. https://doi.org/10.1044/2016_JSLHR-S-15-0296

Snijders, J. T., Tellegen, P. J., Winkel, M., & Laros, J. A. (2003). *Snijders-Oomen Nonverbal Intelligence Test–Revised SON-R 5–17 jaar*. Lisse, the Netherlands: Swets & Zeitlinger.

Sosa, A. V. (2015). Intraword variability in typical speech development. *American Journal of Speech-Language Pathology, 24*(1), 24–35. https://doi.org/10.1044/2014_AJSLP-13-0148

Sosa, A. V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech, Language, and Hearing Research, 55*(2), 596–608. https://doi.org/10.1044/1092-4388(2011/10-0113)

Staiger, A., Schölderle, T., Brendel, B., Bötzel, K., & Ziegler, W. (2017). Oral motor abilities are task dependent: A factor analytic approach to performance rate. *Journal of Motor Behavior, 49*(5), 482–493. https://doi.org/10.1080/00222895.2016.1241747

Terband, H., & Maassen, B. (2012). Childhood apraxia of speech: From symptom based diagnosis to process-oriented diagnosis. *Logopedie en Foniatrie, 84,* 229–234.

Thoonen, G., Maassen, B., Gabreëls, F., & Schreuder, R. (1994). Feature analysis of singleton consonant errors in developmental verbal dyspraxia (DVD). *Journal of Speech and Hearing Research, 37*(6), 1424–1440. https://doi.org/10.1044/jshr.3706.1424

Thoonen, G., Maassen, B., Gabreels, F., & Schreuder, R. (1999). Validity of maximum performance tasks to diagnose motor speech disorders in children. *Clinical Linguistics & Phonetics, 13*(1), 1–23. https://doi.org/10.1080/026992099299211

Thoonen, G., Maassen, B., Wit, J., Gabreëls, F., & Schreuder, R. (1996). The integrated use of maximum performance tasks in differential diagnostic evaluations among children with motor speech disorders. *Clinical Linguistics & Phonetics, 10*(4), 311–336. https://doi.org/10.3109/02699209608985178

van Haaften, L., Diepeveen, S., Van den Engel-Hoek, L., Jonker, M., de Swart, B., & Maassen, B. (2019). The psychometric evaluation of a speech production test battery for children: The reliability and validity of the computer articulation instrument (CAI). *Journal of Speech, Language, and Hearing Research*. Advance online publication. https://doi.org/10.1044/2018_JSLHR-S-18-0274

Van Weerdenburg, M., Verhoeven, L., & Van Balkom, H. (2006). Towards a typology of specific language impairment. *The Journal of Child Psychology and Psychiatry, 47*(2), 176–189. https://doi.org/10.1111/j.1469-7610.2005.01454.x

Vance, M., Stackhouse, J., & Wells, B. (2005). Speech-production skills in children aged 3–7 years. *International Journal of Language & Communication Disorders, 40*(1), 29–48. https://doi.org/10.1080/13682820410001716172

Waring, R., & Knight, R. (2013). How should children with speech sound disorders be classified? A review and critical evaluation of current classification systems. *International Journal of Language & Communication Disorders, 48*(1), 25–40. https://doi.org/10.1111/j.1460-6984.2012.00195.x